一、填空题

1、 Scala语言和Java语言类似，只允许继承（ ）父类。 (易)

参考答案:

【⼀个】

1. 在集群中提交Spark作业时，用于指定每个Executor使用的CPU核⼼数为1个的参数是（ ）。 (易)

参考答案:

【--total-executor-cores 1】

3、 在Spark的TaskScheduler阶段，是通过（ ）管理Task。 (易)

参考答案:

【TaskSetManager】

4、 通常，机器学习的学习形式分类有（ ）和无监督学习。 (易)

参考答案:

【有监督学习】

5、 在Scala中，使用（ ）关键字来定义的类被称为样例类。 (易)

参考答案:

【case】

6、 （ ）表示行键，每个HBase表中只能有⼀个行键，它在HBase中以字典序的方式存储。 (中)

参考答案: 【RowKey】

7、 使用drop命令时，需要先使用（ ）将数据表变为禁用状态。 (中)

参考答案: 【disable】

8、 Spark兼容（ ）、Hive，可以很好地与Hadoop系统融合，从而弥补MapReduce高延迟的性能缺点。 (易)

参考答案: 【HDFS】

9、 SaveMode的参数分别有（ ）、“Overwrite”、（ ）、“Ignore”四个值。 (中)

参考答案:

【Append】【ErrorIfExists】

10、 操作HBase常用的方式有两种，⼀种是（ ），另⼀种是Java API。 (中)

参考答案: 【Shell命令行】

11、 Scala语言可以运行在Windows、（ ）、Mac OS等系统上。(易)

参考答案:【Linux】

12、 在Spark的DAGScheduler阶段，每个（ ）都是TaskSet任务集合。 (中)

参考答案: 【Stage】

13、 KafkaUtils里面提供了两个创建DStream的方式，⼀种是（ ）方式，另⼀种为KafkaUtils.createDirectStream方式。 (中)

参考答案: 【KafkaUtils.createDstream】

14、 创建Scala⼀共有三种类型，分别是Class、（ ）和Trait。 (中)

参考答案: 【Object】

15、 在Spark-Shell中，执行（ ）命令提交任务。 (中)

参考答案: 【spark-submit】

16、 HBase和Hive进行整合，不能将数据直接插⼊到与HBase关联的Hive表中，需要创建（ ）。 (中)

参考答案: 【中间表】

17、 （ ）是⼀种不代表任何意义的值类型，它的作用类似Java中的void。 (中)

参考答案: 【Unit】

18、 只有对于（Key，Value）的RDD，才会有Partitioner（分区），⾮（

Key，Value）的RDD的Parititioner的值是（ ）。 (中)

参考答案: 【None】

19、 当HBase表查询数据遵循的是（ ）。 (中)

参考答案: 【寻址机制】

20、 （ ）是Spark用来处理结构化数据的⼀个模块。

参考答案: 【Spark SQL】

21、 通过（ ）或SQL处理数据，Spark 优化器（Catalyst）会自动优化，即使我们写的程序或SQL不高效，程序也可以高效的执行。 (较难)

参考答案: 【DataFrame API】

22、 Mllib库的主要数据类型包括（ ）、标注点、（ ）。Mllib库的主要数据类型包括（ ）、标注点、（ ）。

参考答案: 【本地向量】【本地矩阵】 (较难)

23、 当前Spark中实现了两种类型的分区函数，⼀个是基于哈希的（ ），另外⼀个是基于范围的RangePartitioner。 (较难)

参考答案: 【HashPartitioner】

24、 若需要停⽌HBase集群，则执行（ ）命令。 (易)

参考答案: 【stop-hbase.sh】

1. 在数据准备阶段，通常将数据随机分为3个模块，即训练数据模块、（ ）和测试数据模块。 （中）

参考答案: 【验证数据模块】

26、 Yarn模式⼜分为（ ）和Yarn Client模式。 （中）

参考答案: 【Yarn Cluster模式】

27、 在创建DataFrame之前，为了支持RDD转换成DataFrame及后续的SQL操作，需要导⼊spark.implicits.\_包启用（ ）。 （中）

参考答案: 【隐式转换】

28、 在流式计算框架的模型中，通常需要构建数据流的拓扑结构，Kafka流处理框架同样是将输⼊主题->（ ）->输出主题抽象成⼀个DAG拓扑图。 （中）

参考答案: 【自定义处理器】

29、 在数据处理应用中，大数据工程师将Spark技术应用于广告、（ ）、推荐系统等业务中。 （中）

参考答案: 【报表】

30、 （ ）对Executor进行状态监控，如果监控到Executor失败，则会⽴刻重新创建。 （中）

参考答案: 【Cluster Manager】

31、 Spark SQL复用了Hive提供的（ ）、HiveQL、（ ）以及序列化和反序列工具（SerDes）。 （中）

参考答案: 【元数据仓库（Metastore）】【用户自定义函数（UDF）】

32、 Scala类抽象机制的扩展通过两种途径实现，⼀种是（ ），另⼀种是混⼊机制，这两种途径都能够避免多重继承的问题。 （较难）

参考答案: 【子类继承】

33、 根据RDD之间依赖关系的不同可以将DAG划分成不同的（ ）。 （较难）

参考答案: 【Stage(调度阶段)】

34、 当Case类不能提前定义的时候，就需要采用（ ）来实现RDD转换成DataFrame。 （较难）

参考答案: 【编程方式定义Schema信息】

35、 Scala旨在与流行的（ ）进行良好的交互操作。 （中）

参考答案: 【Java Runtime Environment（JRE）】

36、 在Spark 1.3.0版本之前，DataFrame被称为（ ）。 （中）

参考答案: 【SchemaRDD】

37、 按照（ ）的理念，Spark在进行任务调度的时候，会尽可能地将计算任务分配到其所要处理数据块的存储位置。 （中）

参考答案: 【移动数据不如移动计算】

38、 Spark为RDD提供了两个重要的机制，分别是持久化机制（即缓存机制）和（ ）。 （中）

参考答案: 【容错机制】

39、 在Scala中，Map是⼀种可迭代的（ ）结构，并且键是唯⼀的，值不⼀定是唯⼀的，所有的值都是通过键来获取的。 （中）

参考答案: 【键值对（key/value）】

40、 在Spark安装目录下，执行（ ）命令可以启动Spark集群。 （易）

参考答案: 【sbin/start-all.sh】

41、 （ ）是指将Spark程序提交至Spark集群中执行任务，由Spark集群负责资源的调度，程序会被框架分发到集群中的节点上并发地执行。 （中）

参考答案: 【集群模式】

42、 MLlib支持的本地矩阵包含（ ）、稀疏矩阵两种类型。 （中）

参考答案: 【密集矩阵】

43、 Hadoop的MapReduce在计算数据时，计算过程必须要转化为（ ）和Reduce两个过程。 （中）

参考答案: 【Map】

44、 测试数据集是为了验证最佳模型，避免（ ）问题。 （中）

参考答案: 【过拟合】

45、

RDD提供了两种故障恢复的方式，分别是（ ）和设置检查点（checkpoint）方式。 （难）

参考答案:

【⾎统（Lineage）方式】

46、 Hive采用MySQL数据库存放Hive元数据，为了能够让Spark访问Hive，就需要将（ ）拷贝到Spark安装路径下的Jars目录。 （中）

参考答案: 【MySQL驱动包】

47、 Scala中的变量在声明时，必须进行（ ）。 （易）

参考答案: 【初始化】

48、 在Scala中，（ ）可以是简写成a.+(b)。 （中）

参考答案: 【a+b】

49、 Spark Streaming支持多种数据源，例如（ ）、Flume以及TCP套接字等数据源。 （中）

参考答案: 【Kafka】

50、 Spark SQL可以通过（ ）从关系型数据库中读取数据的方式创建DataFrame。 （中）

参考答案: 【JDBC】

51、 Spark SQL支持多种数据源的查询和加载，兼容（ ），并可以使用JDBC/ODBC的连接方式来执行SQL语句。 （较难）

参考答案: 【HQL】

52、在DataFrame的操作中，（ ）用于对记录进行分组。 （较难）

参考答案: 【groupBy()】

53、 Spark在2013年加⼊（ ），之后获得迅猛的发展，并于2014年正式成为Apache软件基⾦会的顶级项目。 （较难）

参考答案: 【Apache孵化器项目】

54、 （ ）是Apache Kafka开源项目的⼀个流处理框架。 （较难）

参考答案: 【Kafka Streams】

55、 ⼀个TaskScheduler只能服务⼀个（ ）对象。 （较难）

参考答案: 【SparkContext】

56、 查看所有主题的命令参数是（ ）。 （较难）

参考答案: 【--list】

57、 在Scala中，没有静态方法或静态字段，所以不能用类名直接访问类中的方法和字段，而是创建类的（ ）去访问类中的方法和字段。 （较难）

参考答案: 【实例对象】

58、 Mesos模式下的默认的分区数是（ ）。 （较难）

参考答案: 【8】

59、 在Scala中，控制结构语句包括（ ）和循环语句。 （中）

参考答案: 【条件分支语句】

60、 DataFrame提供了⼀个（ ）以方便操作结构化数据。 （中）

参考答案: 【领域特定语言（DSL）】

61、 （ ）表示时间戳，记录每次操作数据的时间，通常记作数据的版本号。 （中）

参考答案: 【Timestamp】

62、 Spark Streaming的特点有易用性、（ ）、易整合性。 （中）

参考答案: 【容错性】

63、 Spark SQL可以无缝地将（ ）查询与Spark程序进行结合。 （中）

参考答案: 【SQL】

64、 （ ）是从Spark1.6 Alpha版本中引⼊的⼀个新的数据抽象结构，最终在Spark2.0版本被定义成Spark新特性。（中）

参考答案: 【Dataset】

65、 进⼊Windows的命令行，输⼊（ ）命令可以测试Scala环境是否安装成功。 （易）

参考答案:【scala】

66、 Catalyst优化器是⼀个新的可扩展的查询优化器，它是基于（ ）函数式编程结构。 （中）

参考答案:

【Scala】

67、 Spark⽣态系统主要包含（ ）、Spark SQL、（ ）、MLlib、GraphX以及独⽴调度器。 （中）

参考答案:

【Spark Core】【Spark Streaming】

68、 DAG（Directed Acyclic Graph）叫做有向无环图，Spark中的RDD通过⼀系列的转换算子操作和行动算子操作形成了⼀个（ ）。 （中）

参考答案: 【DAG】

69、 在Spark中，不同的RDD之间具有依赖的关系。RDD与它所依赖的RDD的依赖关系有两种类型，分别是（ ）和宽依赖（wide dependency）。 （中）

参考答案: 【窄依赖（narrow dependency）】

70、 （ ）命令用于扫描表并返回表的所有数据。 （中）

参考答案: 【scan】

71、 Scala中的Trait可以被类和对象(Objects)使用关键字（ ）来继承。 （中）

参考答案: 【extends】

72、 Spark运行架构主要是由SparkContext、（ ）和Worker。（难）

参考答案: 【ClusterManager】

73、 在HBase集群中，（ ）负责为HRegion Server分配HRegion。（中）

参考答案: 【HMaster】

74、 Kafka集群是由（ ）、消息代理服务器（

Broker Server）、（ ）组成。 （中）

参考答案: 【⽣产者（Producer）】【消费者（Consumer）】

75、 Spark Streaming提供了⼀个高级抽象的流，即（ ）。

参考答案: 【DStream（离散流）】 （中）

76、 （ ）主要用于返回数据集中的元素个数。（中）

参考答案: 【count()】

77、 通过浏览器访问（ ），查看HBase集群状态。 （中）

参考答案: 【http://node01:16010】

78、 （ ）主要用于以数组的形式返回数据集中的所有元素。 （中）

参考答案: 【collect()】

79、 MLlib库中包含了⼀些通用的机器学习算法和工具类，包括分类、（ ）、聚类、（ ）等。 （中）

参考答案: 【回归】【降维】

80、 MLlib提供了两个分层抽样方法，分别为sampleByKey()、（ ）。（难）

参考答案: 【sampleByKeyExtra()】

81、 Spark Streaming支持从多种数据源获取数据，包括（ ）、Flume、Twitter、ZeroMQ、（ ）、TCP Sockets数据源。 （难）

参考答案: 【Kafka】【Kinesis】

82、 Spark框架对大数据的支持从（ ）、实时处理到交互式查询，进而发展到图计算和（ ）。 （难）

参考答案: 【内存计算】【机器学习模块】

83、 本地矩阵具有整型的（ ）以及Double类型的元素值，它存储在单个机器上。 （难）

参考答案: 【行和列索引值】

84、 Spark要想很好地支持SQL，就需要完成解析（Parser）、（ ）、执行（Execution）三大过程。 （难）

参考答案: 【优化（Optimizer）】

85、 Kafka核⼼API主要有Producer API、（ ）、Streams API、（ ）、AdminClient API五种。 （难）

参考答案: 【Consumer API】【Connect API】

86、 每个（ ）中保存的是HBase表中某段连续的数据。 （难）

参考答案: 【Region】

87、 （ ）是⼀种以RDD为基础的分布式数据集，因此它可以完成RDD的绝大多数功能。 （中）

参考答案: 【DataFrame】

88、 （ ）是Spark提供的可扩展的机器学习库，其特点是采用较为先进的迭代式、内存存储的分析计算，使得数据的计算处理速度大大高于普通的数据处理引擎。 （中）

参考答案: 【MLlib】

89、 在协同过滤算法中有着两个分支，分别是基于群体用户的协同过滤（UserCF）、（ ）。 （中）

参考答案: 【基于物品的协同过滤（ItemCF）】

90、 Kafka的设计初衷是为实时数据提供⼀个（ ）、高通量、（ ）的消息传递平台。 （中）

参考答案: 【统⼀】【低等待】

91、 Spark SQL的重要特点是能够统⼀处理（ ）和RDD。 （较难）

参考答案:【关系表】

92、 （ ）操作会筛选出满⾜函数func的元素，并返回⼀个新的数据集。 （较难）

参考答案: 【filter(func)】

93、 Scala中的数组分为（ ）和变长数组。 （易）

参考答案: 【定长数组】

94、 Spark的任务调度流程分为RDD Objects、（ ）、（ ）以及Worker四个部分。 （较难）

参考答案: 【DAGScheduler】【TaskScheduler】

95、 Spark可以访问各种数据源，包括本地⽂件系统、（ ）、Cassandra、（ ）和Hive等。 （中）

参考答案: 【HDFS】【HBase】

96、 Spark Streaming具有很好的（ ），在没有额外代码和配置的情况下，可以恢复丢失的数据。 （中）

参考答案: 【容错性】

97、 Kafka的消息传递模式有（ ）、发布订阅消息传递模式。 （中）

参考答案: 【点对点消息传递模式】

98、 当join算子操作后，分区数量（ ）则为窄依赖。 （中）

参考答案: 【没有变化】

99、 Kafka集群中消息的消费模型有两种，分别是（ ）和（ ）。 （中）

参考答案: 【推送模型（push）】【拉取模型（pull）】

100、Client通过请求（ ）地址来查询目标数据。 （中）

参考答案: 【RegionServer】